Renibacterium salmoninarum: Genome Sequencing, Finishing Strategies and Assembly Validation

Rajinder Kaul, Ph.D.

University of Washington

Genome Center

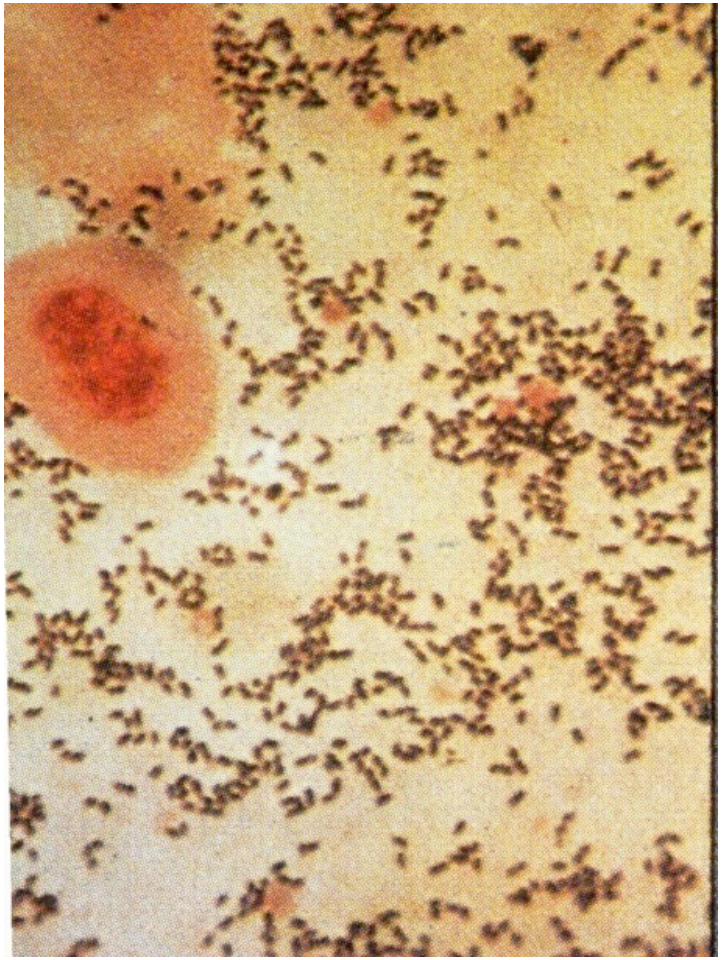
Microbial Genomics Era

- First microbial genome sequenced in 1995 *Haemophillus influenza* RD strain
- Number microbial genomes listed : 827
- Completed: 281
  - archae: 25
  - bacteria: 257
- In Sequencing: 546
  - archae: 25
  - bacteria: 521
- UWGC microbial WGS
  - Completed: 7 organisms
  - Total microbial genome size: 22 Mb
  - In sequencing pipeline: 23 Mb

http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi

Renibacterium salmoninarum

- Gram+ve diplococcobacillus, member of the *Micrococcaceae* family
- Grows slowly *in vitro*
  - optimum temperature 15°C
  - doubling time 24 hours
- Infects salmonids,
  - transmitted horizontally and vertically
  - causes bacterial kidney disease (BKD)
- 3.5 Mb genome



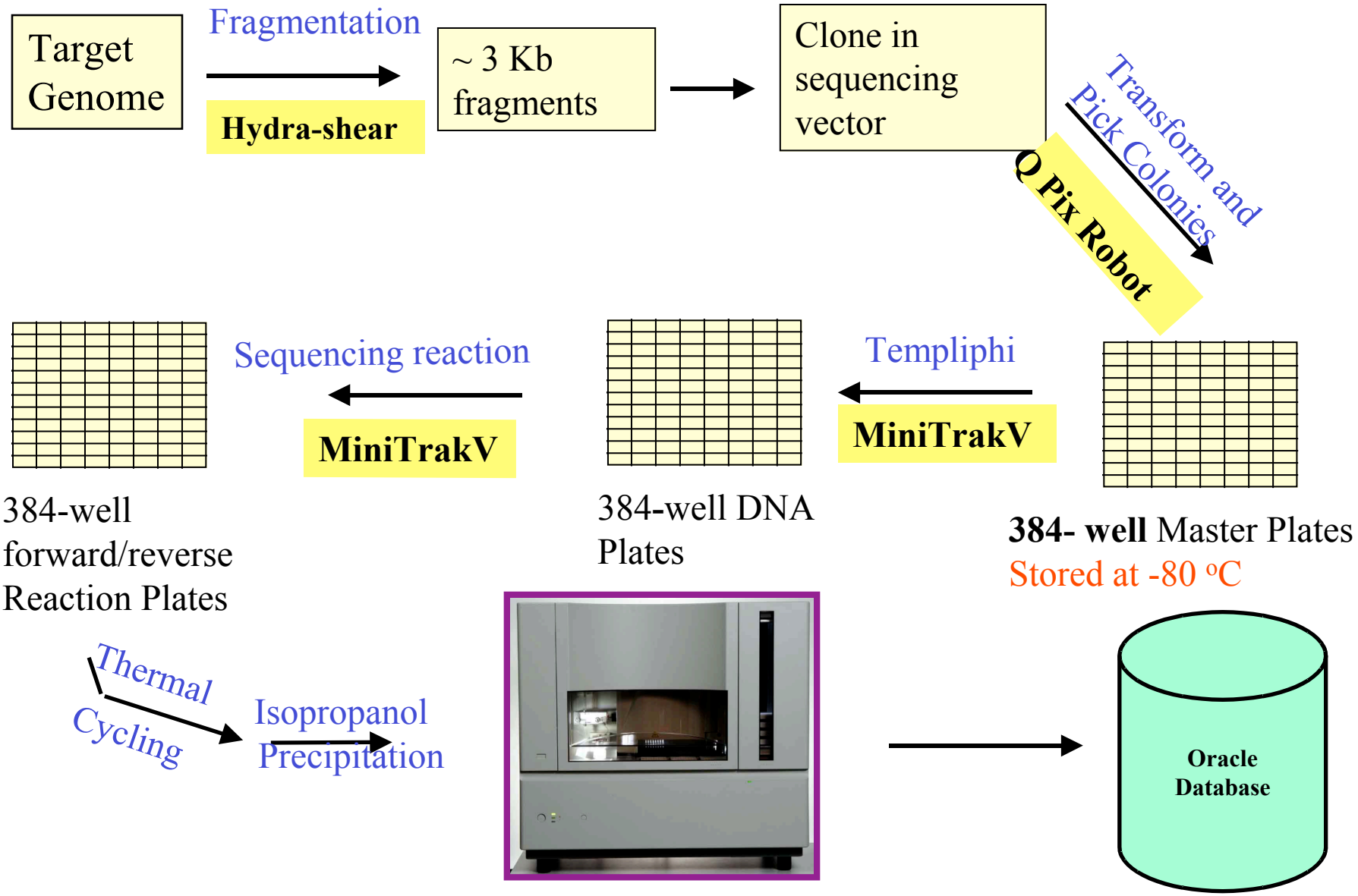
Courtesy: Mark Strom

BKD Incidence

- Predominant salmonid bacterial disease
- 30-85% incidence in hatchery stocks and wild stocks
- Up to 80% mortality in captive stocks
- 15% mortality in captive brood stocks due to BKD since 2000

Courtesy: Mark Strom

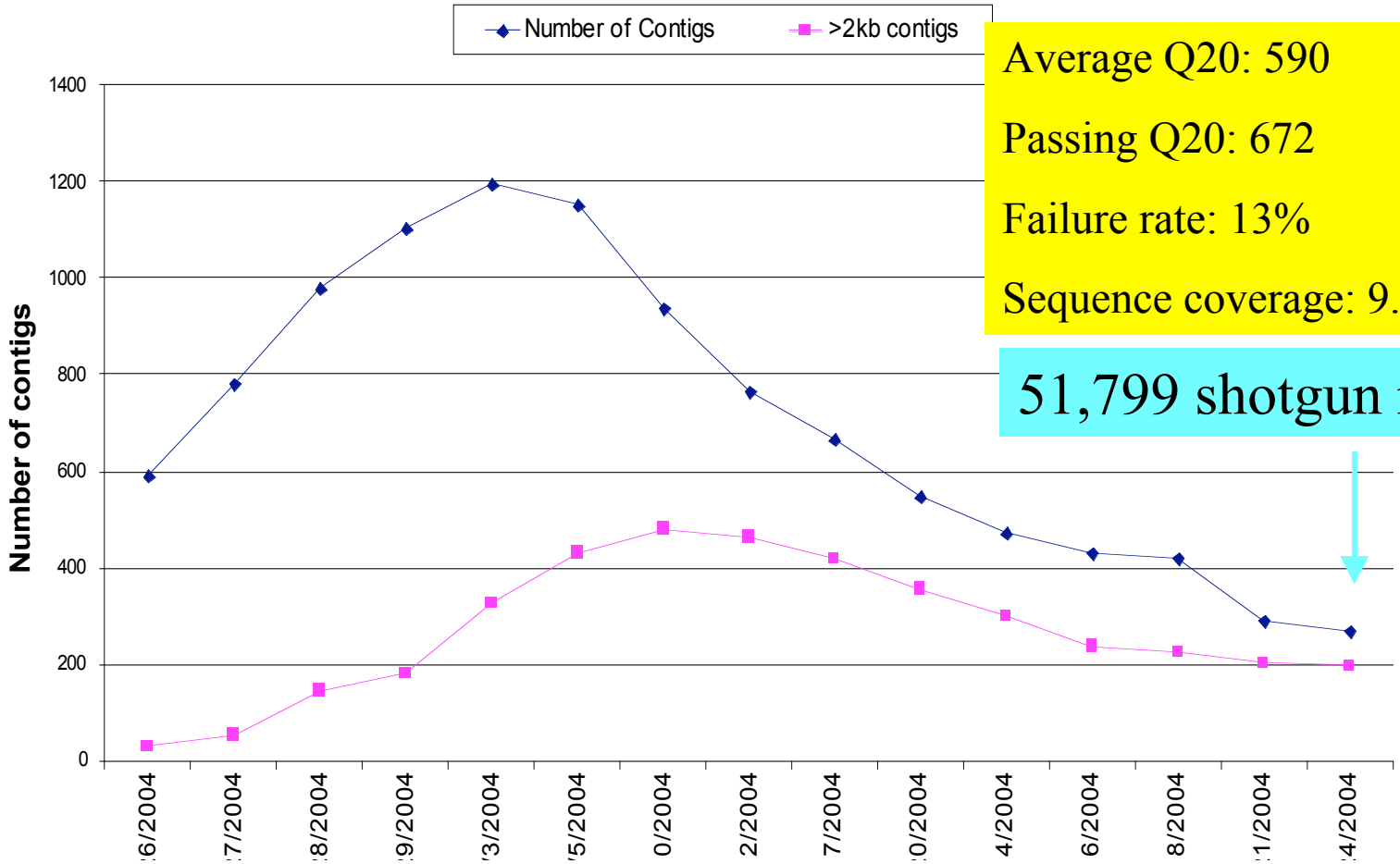
Schematics of Sequencing Pipeline



Some Terminologies

- PHRAP: Software tool for pair-wise alignment of individual reads for assembling in to contiguous chromosome
- PHRED quality: Statistical quality value assignments to trace data
- CONSED: Software tool for viewing genome assemblies and its manual manipulations
- AUTOFINISH: Automated evaluation of sequence quality and contiguity and design experiments to improve quality and attempts to close gaps
- Q20: Logarithmic scale error probability value. Q20 implies probability of 1 in 100 a base could be wrong

Sequencing Progress of Renibacterium salmoninarum



Average Q20: 590

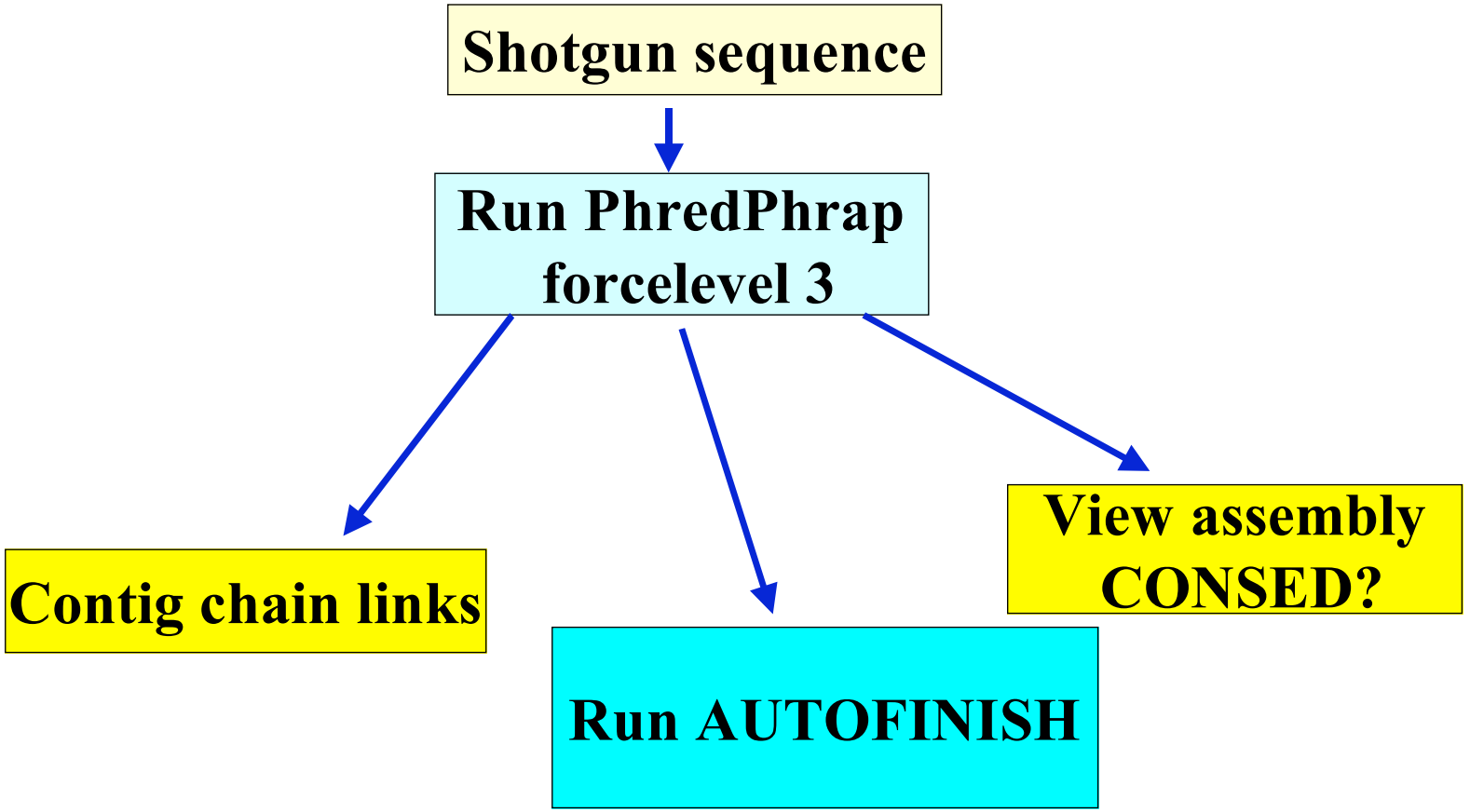
Passing Q20: 672

Failure rate: 13%

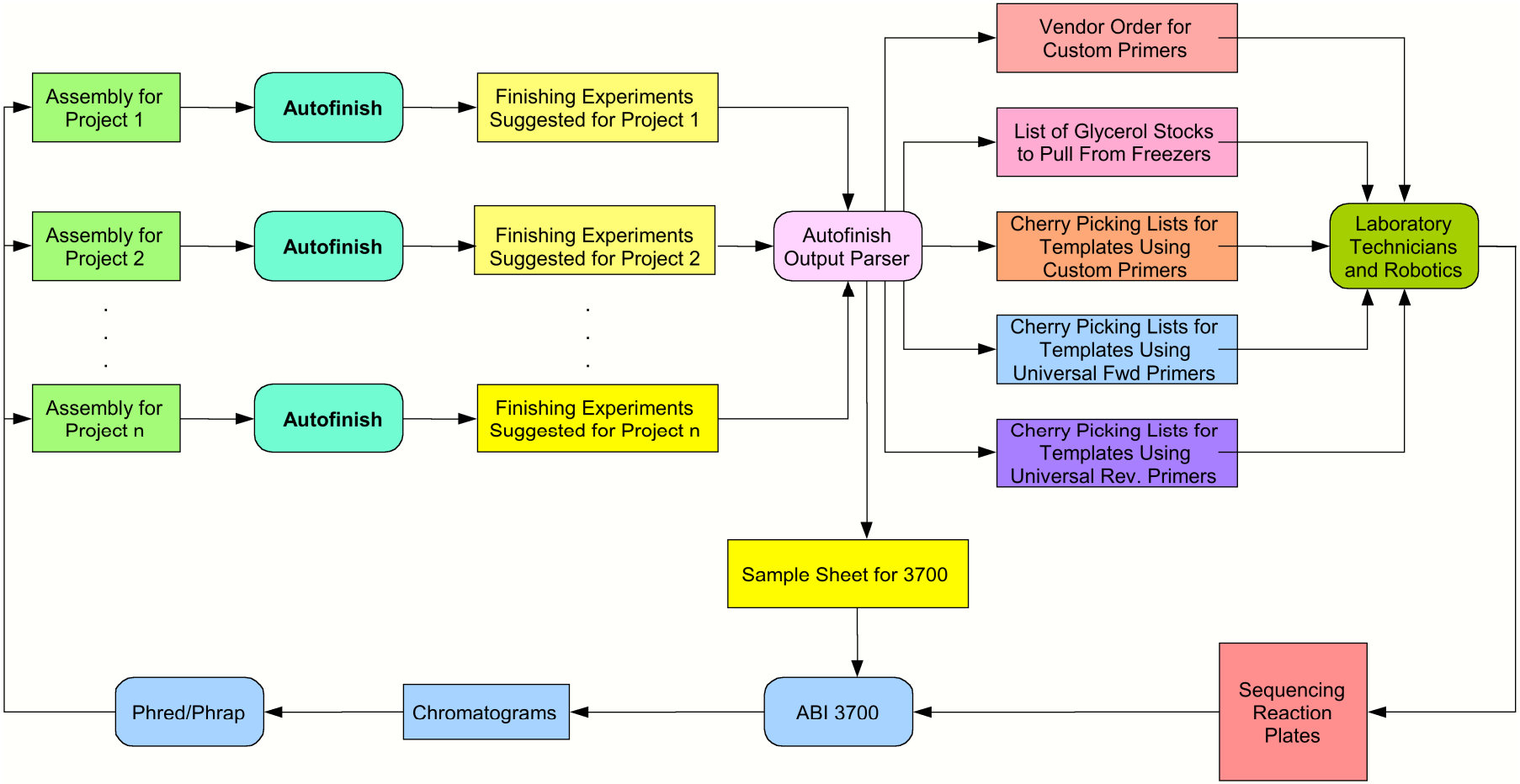
Sequence coverage: 9.7 X

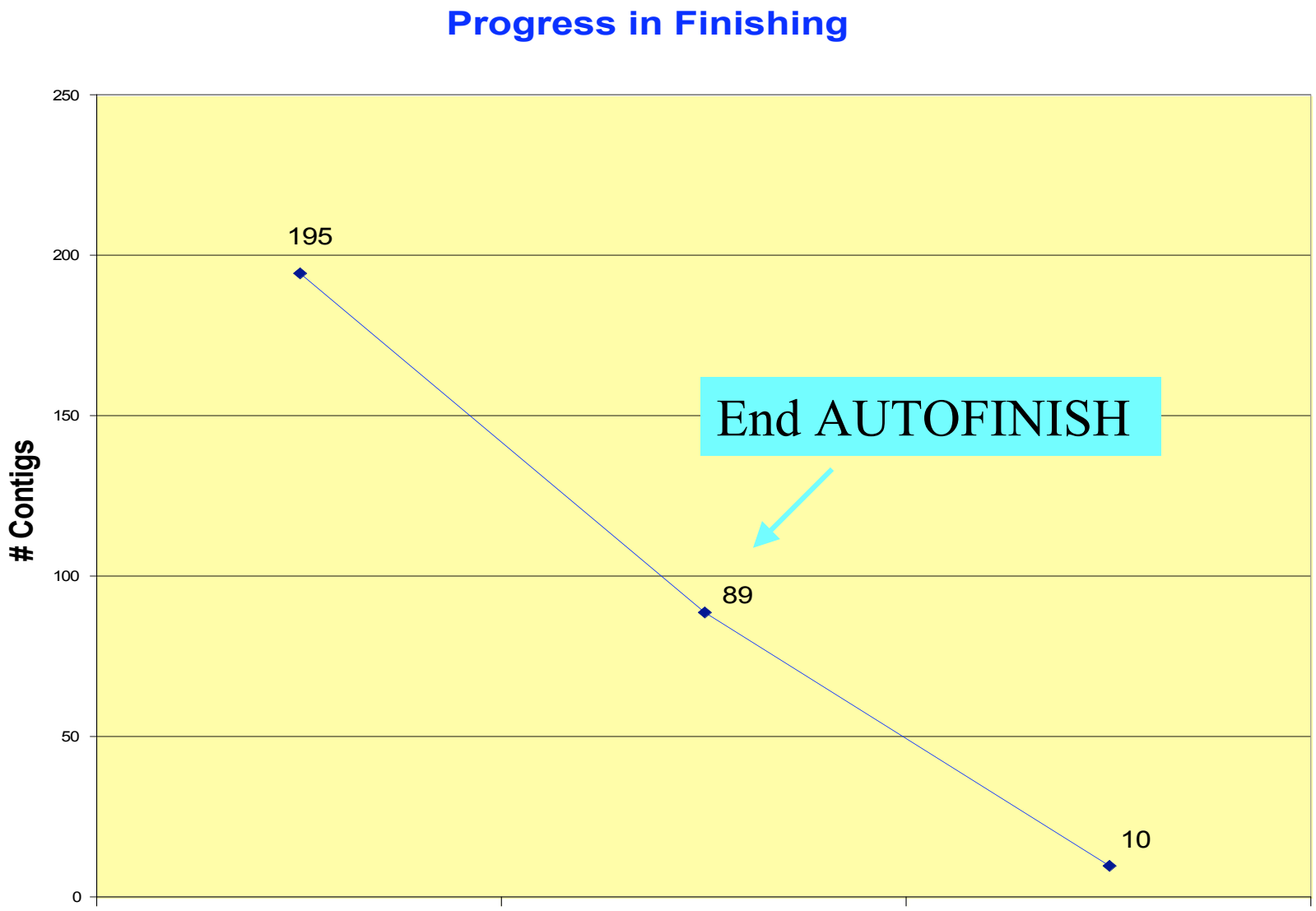
51,799 shotgun reads

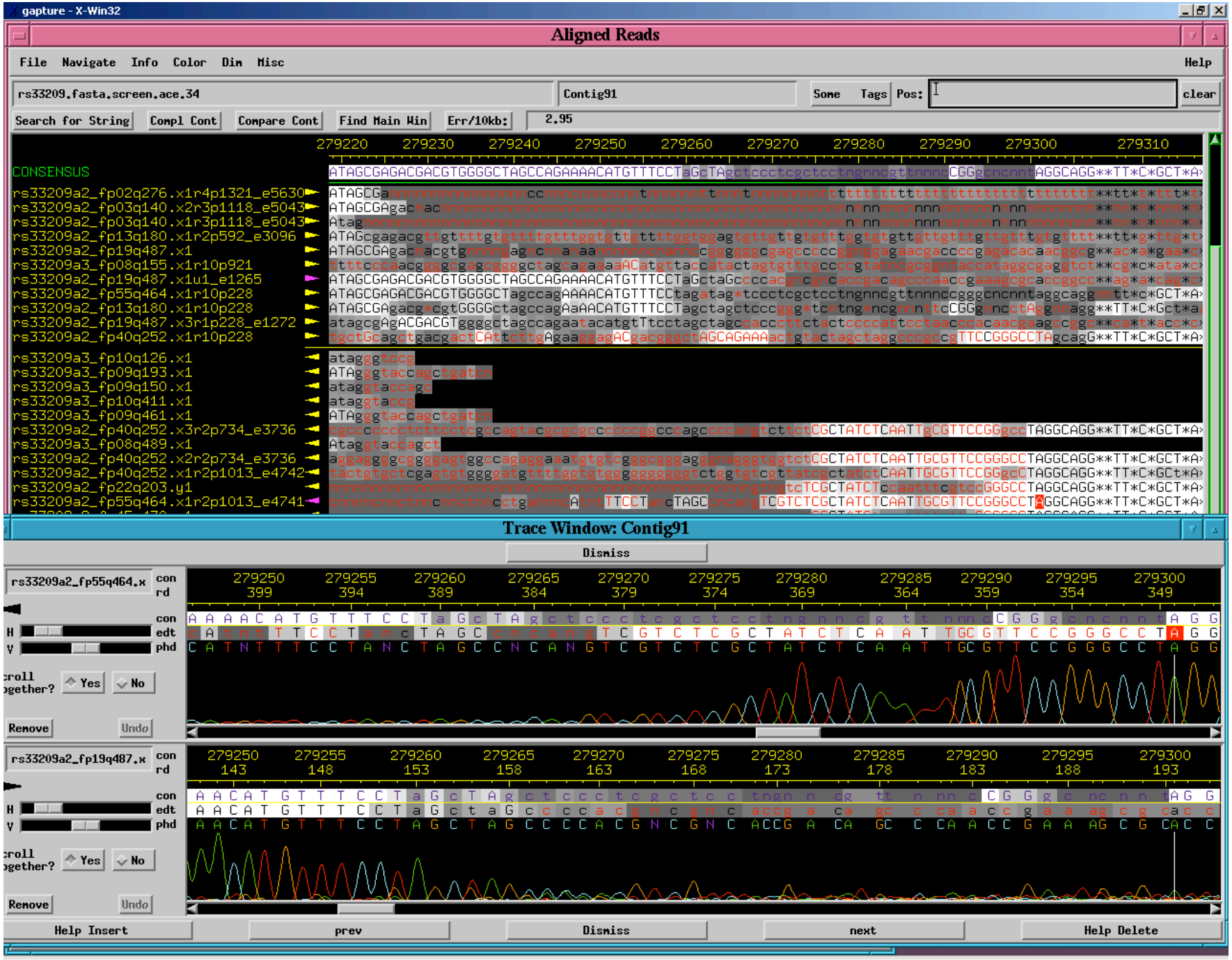
Project Evaluation Following Shotgun Sequencing

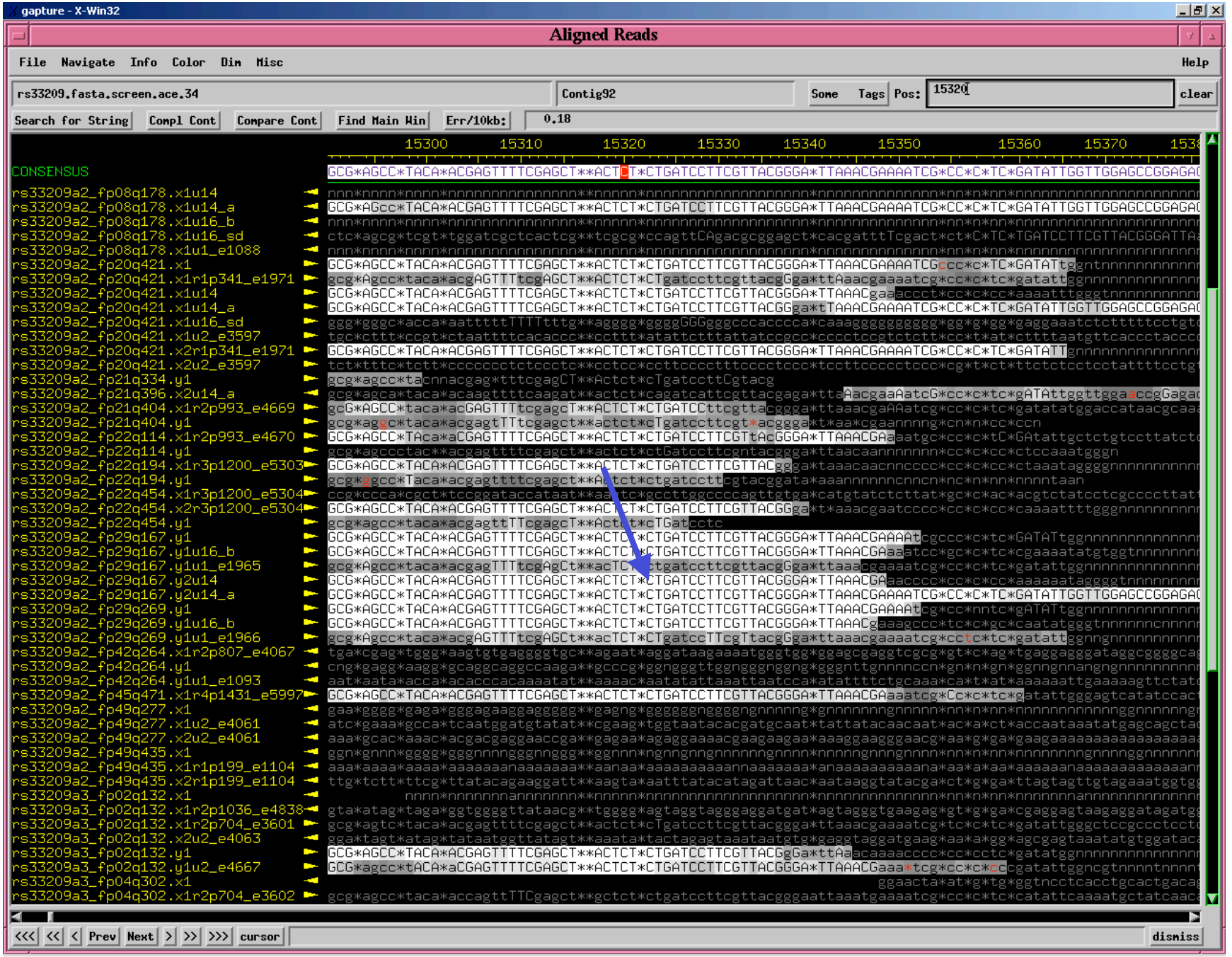


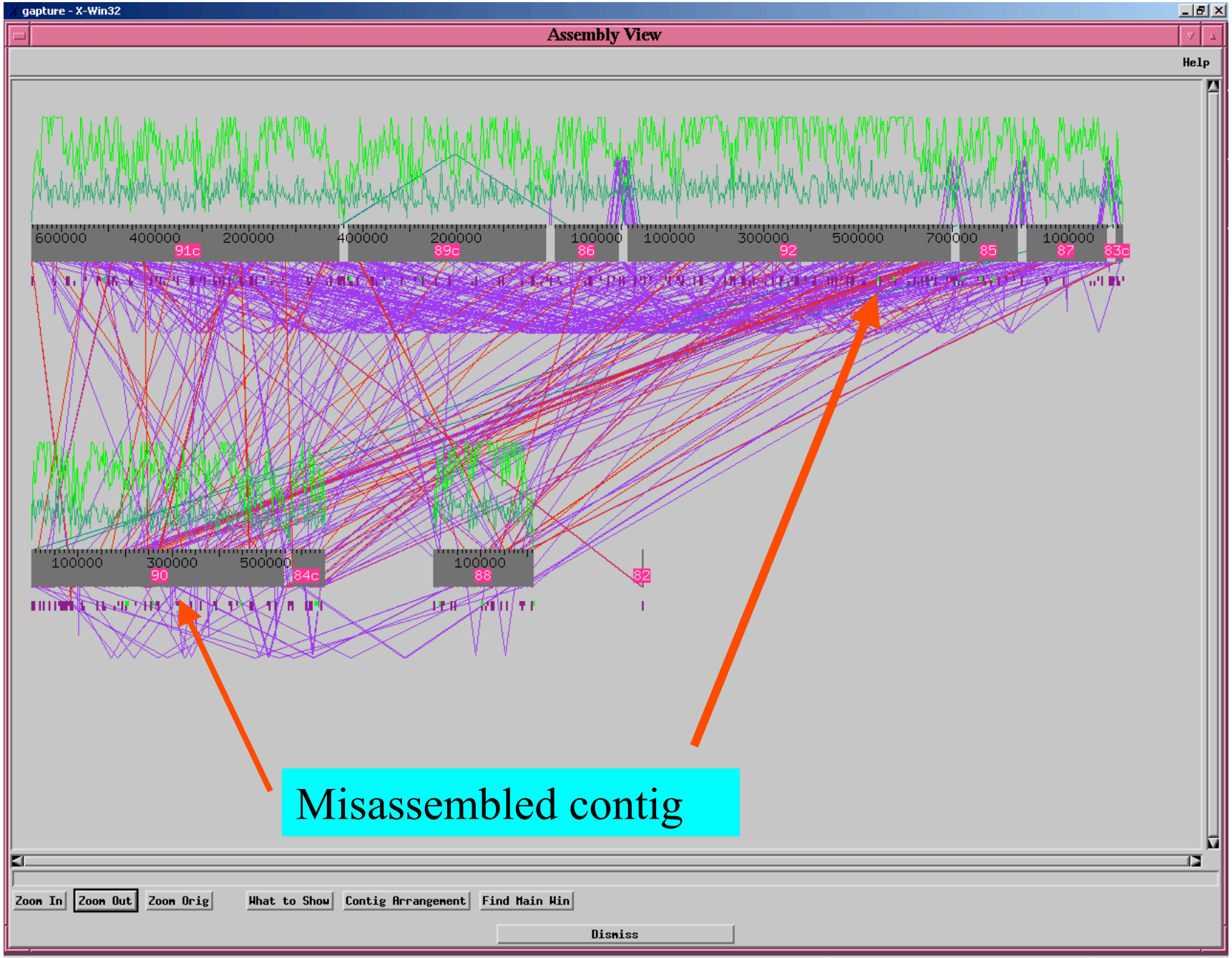
Automated Finishing Dataflow




Progress in Finishing
Contigs
250
200
150
100
50
0
195
89
10
End AUTOFINISH

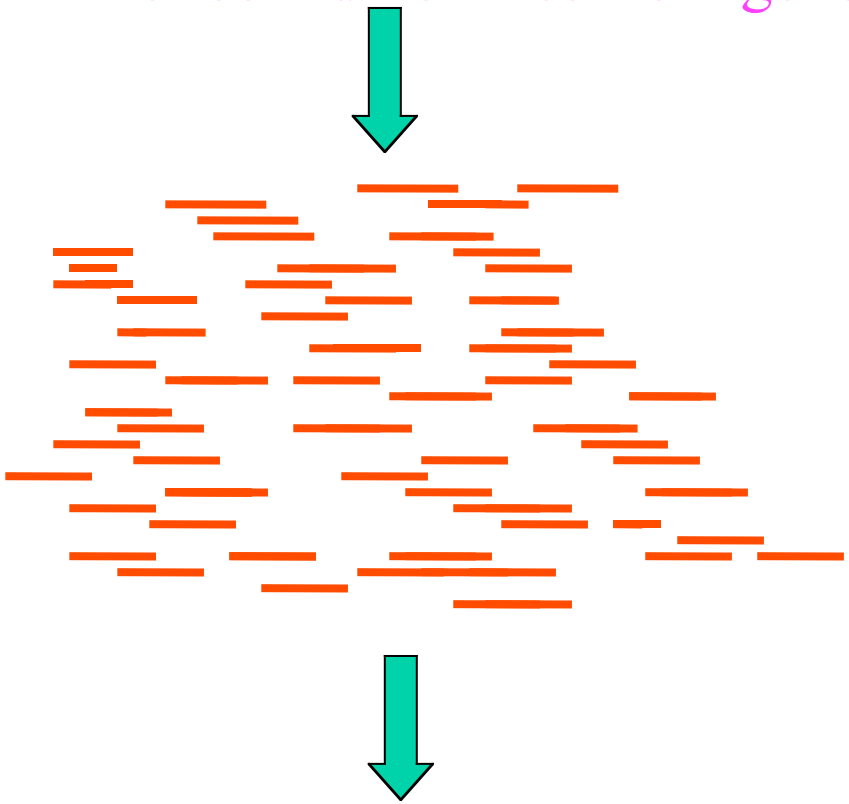

gapture - X-Win32
Aligned Reads
File Navigate Info Color Dim Misc Help
rs33209.fasta.screen.ace.34
Contig91
Some Tags
Pos:
clear
Search for String
Compl Cont
Compare Cont
Find Main Win
Err/10kb:
2.95
279220 279230 279240 279250 279260 279270 279280 279290 279300 279310
CONSENSUS
rs33209a2_fp02q276.x1r4p1321_e5630
rs33209a2_fp03q140.x2r3p1118_e5043
rs33209a2_fp03q140.x1r3p1118_e5043
rs33209a2_fp13q180.x1r2p592_e3096
rs33209a2_fp19q487.x1
rs33209a3_fp08q155.x1r10p921
rs33209a2_fp19q487.x1u1_e1265
rs33209a2_fp55q464.x1r10p228
rs33209a2_fp13q180.x1r10p228
rs33209a2_fp19q487.x3r1p228_e1272
rs33209a2_fp40q252.x1r10p228
rs33209a3_fp10q126.x1
rs33209a3_fp09q193.x1
rs33209a3_fp09q150.x1
rs33209a3_fp10q411.x1
rs33209a3_fp09q461.x1
rs33209a2_fp40q252.x3r2p734_e3736
rs33209a3_fp08q489.x1
rs33209a2_fp40q252.x2r2p734_e3736
rs33209a2_fp40q252.x1r2p1013_e4742
rs33209a2_fp22q203.y1
rs33209a2_fp55q464.x1r2p1013_e4741
Trace Window: Contig91
Dismiss
rs33209a2_fp55q464.x
con
rd
279250 279255 279260 279265 279270 279275 279280 279285 279290 279295 279300
399 394 389 384 379 374 369 364 359 354 349
H
V
edt
phd
Scroll together? Yes No
Remove
Undo
rs33209a2_fp19q487.x
279250 279255 279260 279265 279270 279275 279280 279285 279290 279295 279300
143 148 153 158 163 168 173 178 183 188 193
Help Insert
prev
Dismiss
next
Help Delete


gapture - X-Win32
Aligned Reads
File Navigate Info Color Dim Misc Help
rs33209.fasta.screen.ace.34
Contig92
Some Tags
Pos: 15320
clear
Search for String
Compl Cont
Compare Cont
Find Main Win
Err/10kb:
0.18
15300 15310 15320 15330 15340 15350 15360 15370
CONSENSUS
rs33209a2_fp08q178.x1u14
rs33209a2_fp08q178.x1u14_a
rs33209a2_fp08q178.x1u16_b
rs33209a2_fp08q178.x1u16_sd
rs33209a2_fp08q178.x1u1_e1088
rs33209a2_fp20q421.x1
rs33209a2_fp20q421.x1r1p341_e1971
rs33209a2_fp20q421.x1u14
rs33209a2_fp20q421.x1u14_a
rs33209a2_fp20q421.x1u16_sd
rs33209a2_fp20q421.x1u2_e3597
rs33209a2_fp20q421.x2r1p341_e1971
rs33209a2_fp20q421.x2u2_e3597
rs33209a2_fp21q334.y1
rs33209a2_fp21q396.x2u14_a
rs33209a2_fp21q404.x1r2p993_e4669
rs33209a2_fp21q404.y1
rs33209a2_fp22q114.x1r2p993_e4670
rs33209a2_fp22q114.y1
rs33209a2_fp22q194.x1r3p1200_e5303
rs33209a2_fp22q194.y1
rs33209a2_fp22q454.x1r3p1200_e5304
rs33209a2_fp22q454.x2r3p1200_e5304
rs33209a2_fp22q454.y1
rs33209a2_fp29q167.y1
rs33209a2_fp29q167.y1u16_b
rs33209a2_fp29q167.y1u1_e1965
rs33209a2_fp29q167.y2u14
rs33209a2_fp29q167.y2u14_a
rs33209a2_fp29q269.y1
rs33209a2_fp29q269.y1u16_b
rs33209a2_fp29q269.y1u1_e1966
rs33209a2_fp42q264.x1r2p807_e4067
rs33209a2_fp42q264.y1
rs33209a2_fp42q264.y1u1_e1093
rs33209a2_fp45q471.x1r4p1431_e5997
rs33209a2_fp49q277.x1
rs33209a2_fp49q277.x1u2_e4061
rs33209a2_fp49q277.x2u2_e4061
rs33209a2_fp49q435.x1
rs33209a2_fp49q435.x1r1p199_e1104
rs33209a2_fp49q435.x2r1p199_e1104
rs33209a3_fp02q132.x1
rs33209a3_fp02q132.x1r2p1036_e4838
rs33209a3_fp02q132.x1r2p704_e3601
rs33209a3_fp02q132.x2u2_e4063
rs33209a3_fp02q132.y1
rs33209a3_fp02q132.y1u2_e4667
rs33209a3_fp04q302.x1
rs33209a3_fp04q302.x1r2p704_e3602
<<< << < Prev Next > >> >>> cursor
dismiss

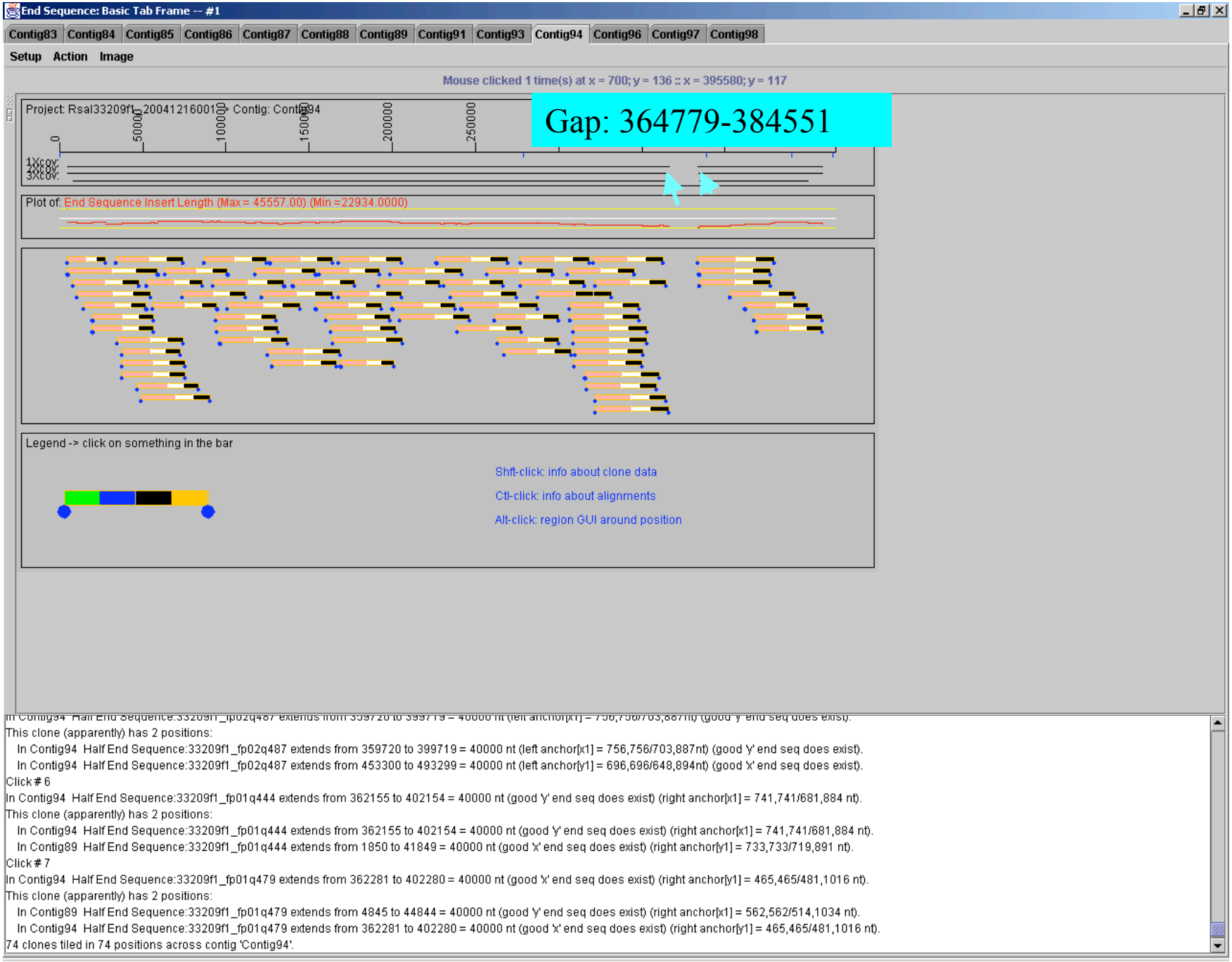

gapture - X-Win32
Assembly View
Help
600000
400000
200000
91c
400000
200000
89c
100000
86
100000
300000
500000
700000
92
85
100000
87
83c
100000
300000
500000
90
84c
100000
88
82
Misassembled contig
Zoom In
Zoom Out
Zoom Orig
What to Show
Contig Arrangement
Find Main Win
Dismiss

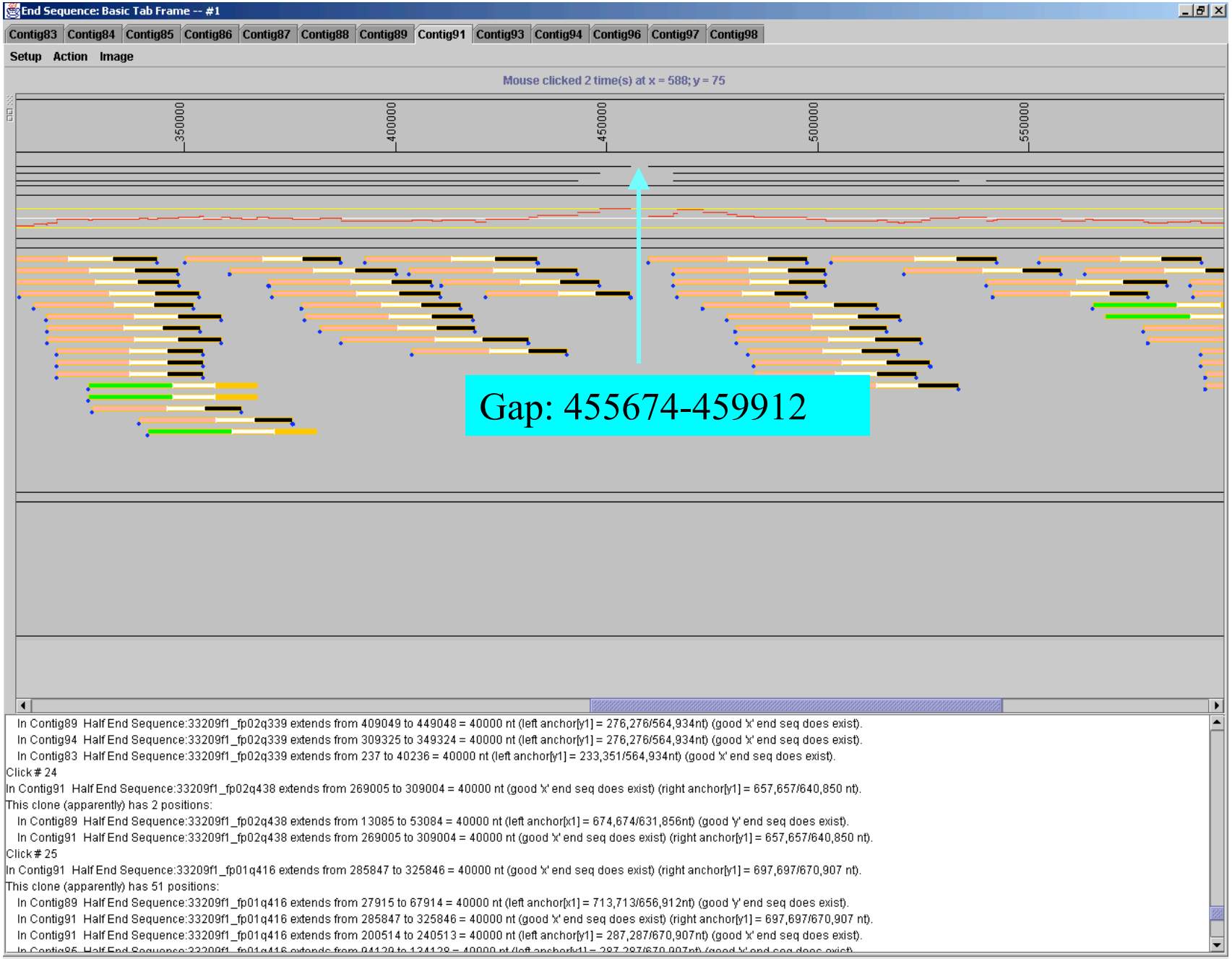
Assembly Validation
Genomic Variation

Make fosmid libraries from genome

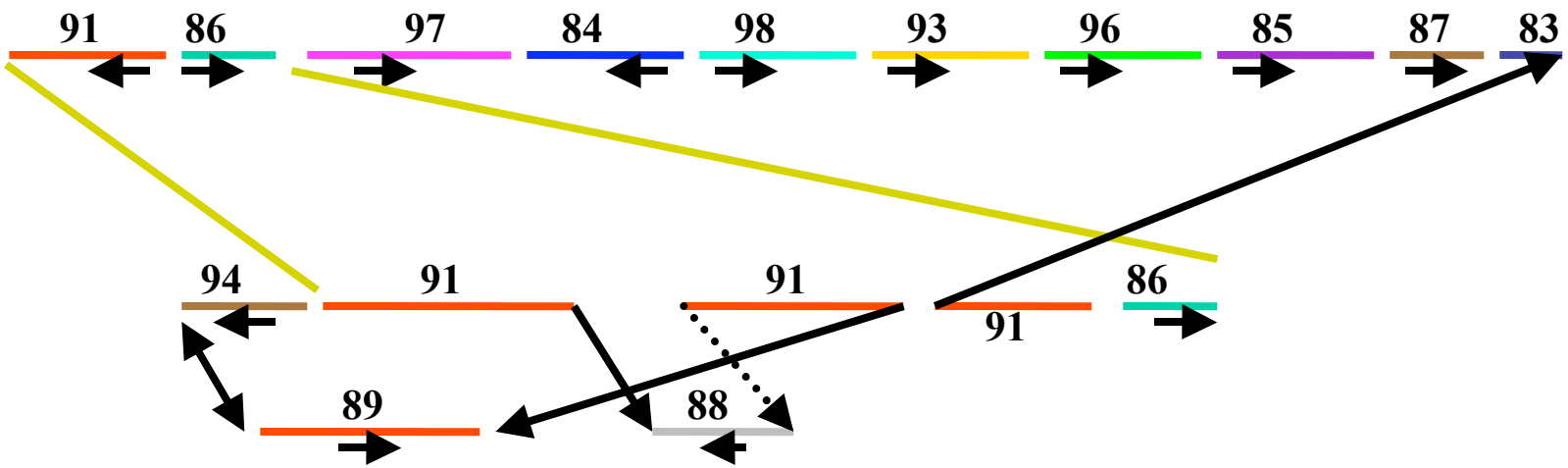


End Sequence and Fingerprint data used to validate genome assembly


End Sequence: Basic Tab Frame -- #1
Contig83 Contig84 Contig85 Contig86 Contig87 Contig88 Contig89 Contig91 Contig93 Contig94 Contig96 Contig97 Contig98
Setup Action Image
Mouse clicked 1 time(s) at x = 700; y = 136 :: x = 395580; y = 117
Project: Rsal33209f1_200412160010 Contig: Contig94
Gap: 364779-384551
0
50000
100000
150000
200000
250000
1Xcov:
2Xcov:
3Xcov:
Plot of: End Sequence Insert Length (Max = 45557.00) (Min = 22934.0000)
Legend -> click on something in the bar
Shft-click: info about clone data
Ctl-click: info about alignments
Alt-click: region GUI around position
This clone (apparently) has 2 positions:
In Contig94 Half End Sequence:33209f1_fp02q487 extends from 359720 to 399719 = 40000 nt (left anchor[x1] = 756,756/703,887nt) (good 'y' end seq does exist).
In Contig94 Half End Sequence:33209f1_fp02q487 extends from 453300 to 493299 = 40000 nt (left anchor[y1] = 696,696/648,894nt) (good 'x' end seq does exist).
Click # 6
In Contig94 Half End Sequence:33209f1_fp01q444 extends from 362155 to 402154 = 40000 nt (good 'y' end seq does exist) (right anchor[x1] = 741,741/681,884 nt).
This clone (apparently) has 2 positions:
In Contig94 Half End Sequence:33209f1_fp01q444 extends from 362155 to 402154 = 40000 nt (good 'y' end seq does exist) (right anchor[x1] = 741,741/681,884 nt).
In Contig89 Half End Sequence:33209f1_fp01q444 extends from 1850 to 41849 = 40000 nt (good 'x' end seq does exist) (right anchor[y1] = 733,733/719,891 nt).
Click # 7
In Contig94 Half End Sequence:33209f1_fp01q479 extends from 362281 to 402280 = 40000 nt (good 'x' end seq does exist) (right anchor[y1] = 465,465/481,1016 nt).
This clone (apparently) has 2 positions:
In Contig89 Half End Sequence:33209f1_fp01q479 extends from 4845 to 44844 = 40000 nt (good 'y' end seq does exist) (right anchor[x1] = 562,562/514,1034 nt).
In Contig94 Half End Sequence:33209f1_fp01q479 extends from 362281 to 402280 = 40000 nt (good 'x' end seq does exist) (right anchor[y1] = 465,465/481,1016 nt).
74 clones tiled in 74 positions across contig 'Contig94'.


End Sequence: Basic Tab Frame -- #1
Contig83 Contig84 Contig85 Contig86 Contig87 Contig88 Contig89 Contig91 Contig93 Contig94 Contig96 Contig97 Contig98
Setup Action Image
Mouse clicked 2 time(s) at x = 588; y = 75
3500000
4000000
4500000
5000000
5500000
Gap: 455674-459912
In Contig89 Half End Sequence:33209f1_fp02q339 extends from 409049 to 449048 = 40000 nt (left anchor[y1] = 276,276/564,934nt) (good 'x' end seq does exist).
In Contig94 Half End Sequence:33209f1_fp02q339 extends from 309325 to 349324 = 40000 nt (left anchor[y1] = 276,276/564,934nt) (good 'x' end seq does exist).
In Contig83 Half End Sequence:33209f1_fp02q339 extends from 237 to 40236 = 40000 nt (left anchor[y1] = 233,351/564,934nt) (good 'x' end seq does exist).
Click # 24
In Contig91 Half End Sequence:33209f1_fp02q438 extends from 269005 to 309004 = 40000 nt (good 'x' end seq does exist) (right anchor[y1] = 657,657/640,850 nt).
This clone (apparently) has 2 positions:
In Contig89 Half End Sequence:33209f1_fp02q438 extends from 13085 to 53084 = 40000 nt (left anchor[x1] = 674,674/631,856nt) (good 'y' end seq does exist).
In Contig91 Half End Sequence:33209f1_fp02q438 extends from 269005 to 309004 = 40000 nt (good 'x' end seq does exist) (right anchor[y1] = 657,657/640,850 nt).
Click # 25
In Contig91 Half End Sequence:33209f1_fp01q416 extends from 285847 to 325846 = 40000 nt (good 'x' end seq does exist) (right anchor[y1] = 697,697/670,907 nt).
This clone (apparently) has 51 positions:
In Contig89 Half End Sequence:33209f1_fp01q416 extends from 27915 to 67914 = 40000 nt (left anchor[x1] = 713,713/656,912nt) (good 'y' end seq does exist).
In Contig91 Half End Sequence:33209f1_fp01q416 extends from 285847 to 325846 = 40000 nt (good 'x' end seq does exist) (right anchor[y1] = 697,697/670,907 nt).
In Contig91 Half End Sequence:33209f1_fp01q416 extends from 200514 to 240513 = 40000 nt (left anchor[y1] = 287,287/670,907nt) (good 'x' end seq does exist).

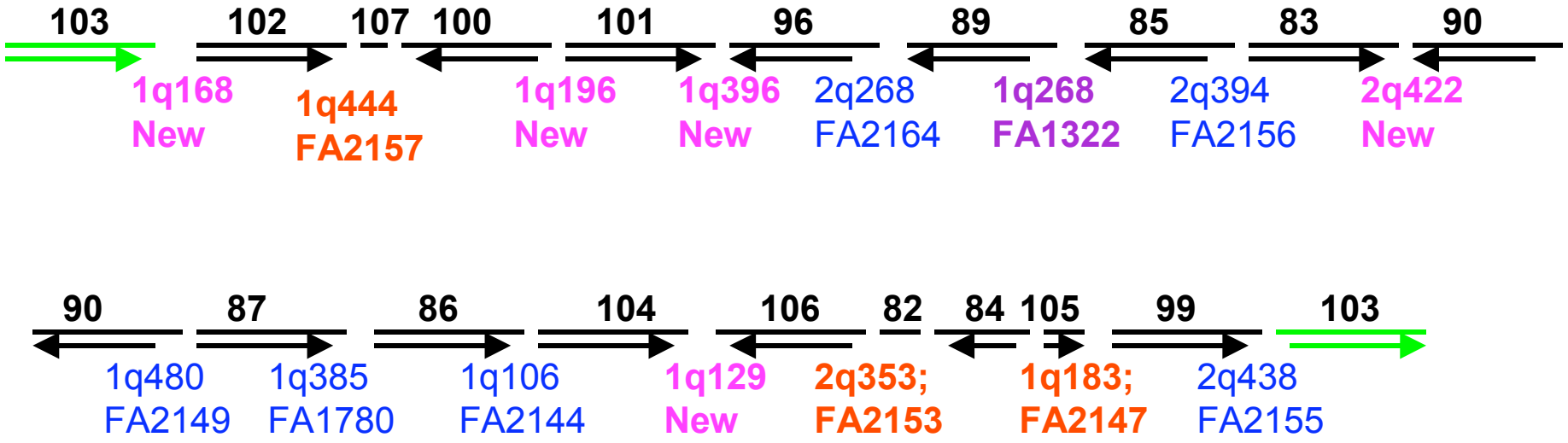
Contig Order and Orientation



- 11 original contigs
- Fosmid end sequence tiled pointing misassembled regions
- Identify fosmids to fix misassemblies

Manual Tearing and Order-Orientation

***Renibacterium salmoninarum* Ace.40; 7/13/2005**



- Tore contigs to create 20 from 11 original contigs
- Individual fosmids were sequenced
- Backbones from fosmid assemblies used to fix misassemblies
- Low quality regions improved by additional sequencing

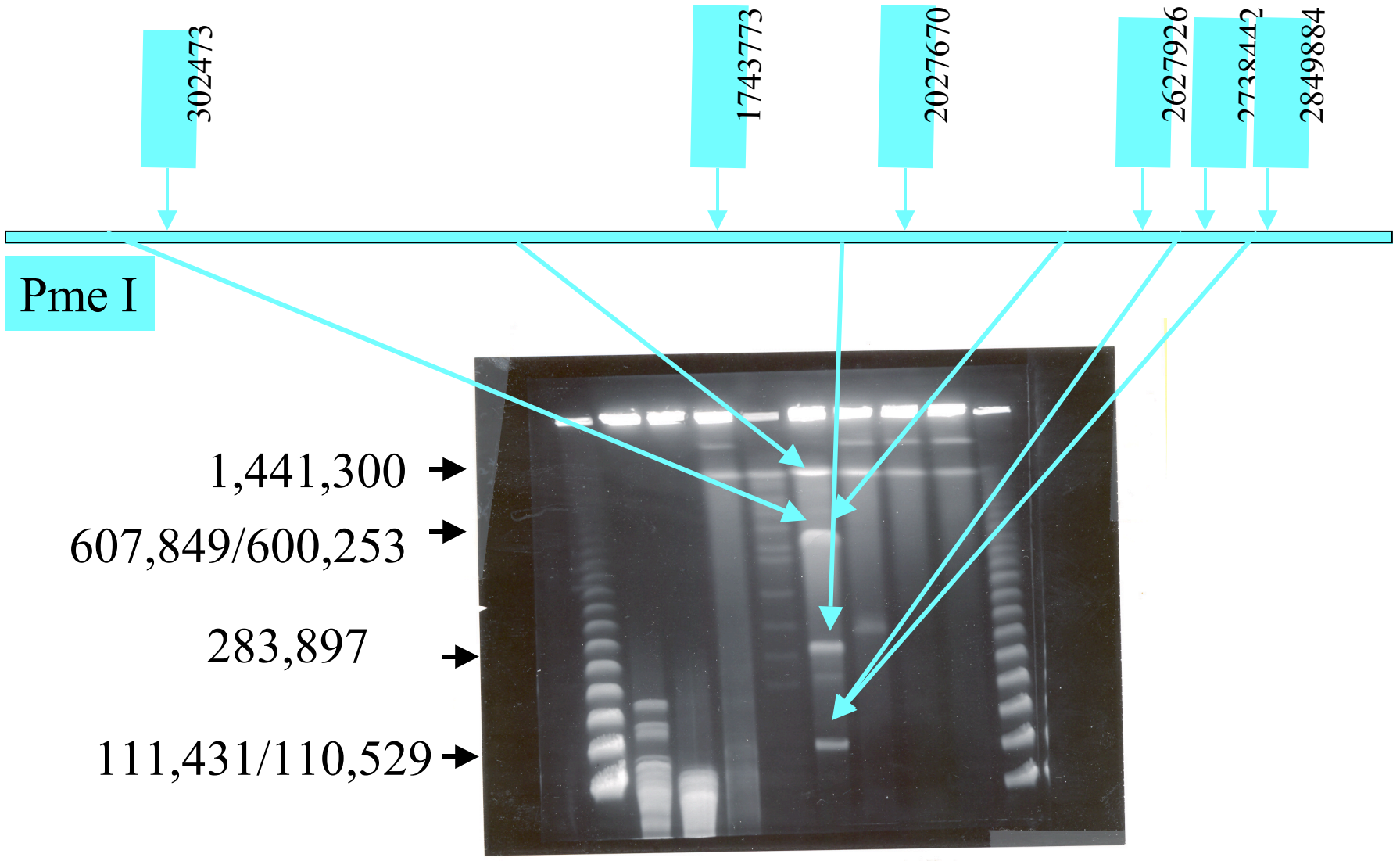
Renibacterium salmoninarum Genome

- Single Circular chromosome with **3,155,258** bases
- 52692 reads in final assembly
- 51799 shotgun reads
- 8968 finishing reads
- 29 fosmid clones sequenced independently to fix misassemblies (~22,272 reads)
- 83,932 total reads

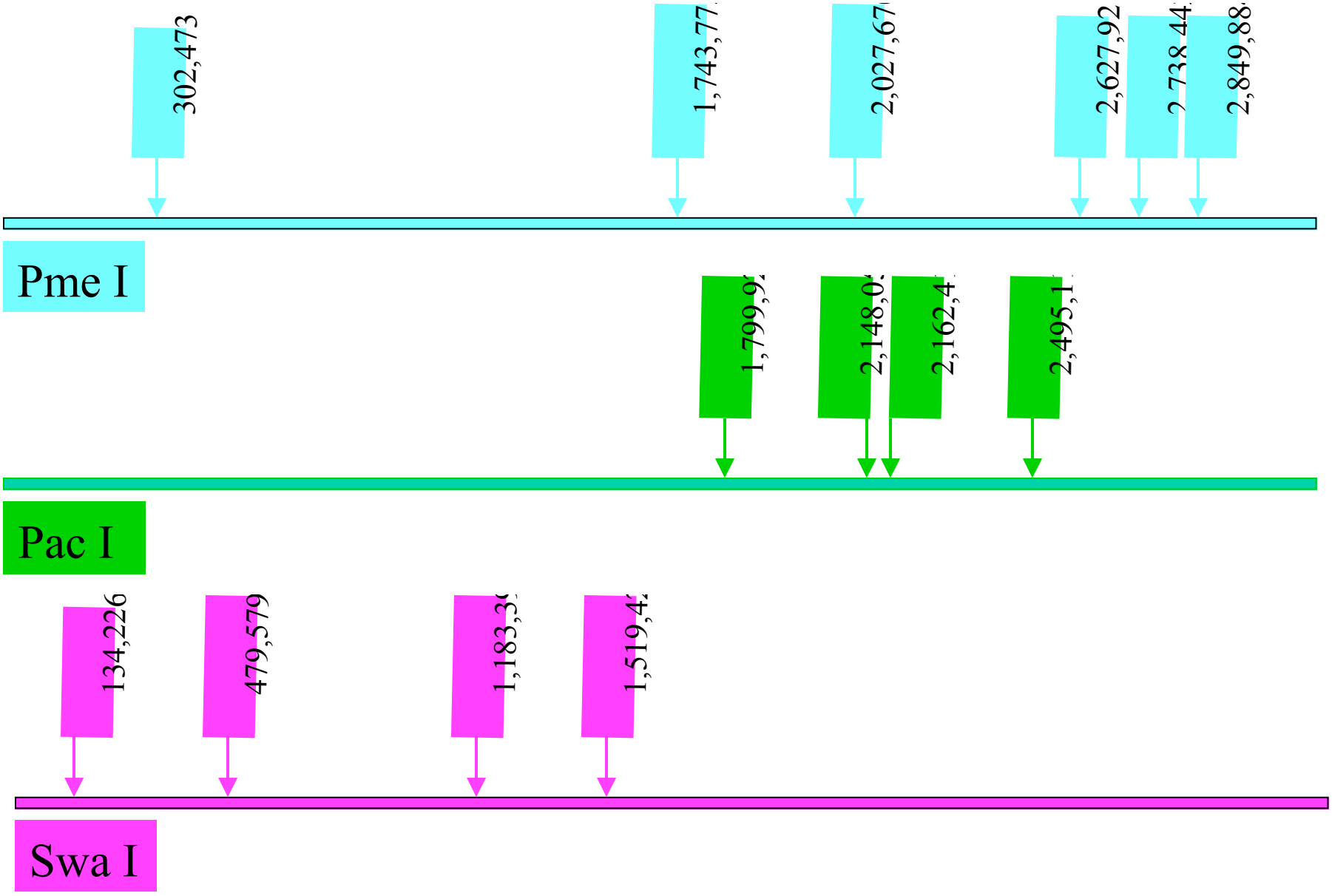
Genome Assembly Validation

- Gross level validation
  - Compare experimental and virtual sequence derived PFGE fragment distribution data derived from rare cutting restriction enzymes
- Finer scale assembly validation (1 Kb resolution)
  - Tile fosmid end-sequences and corresponding fingerprint data in Hind III, Pst I, and Bgl II domains against finished genome assembly
  - All sequence derived virtual fingerprint fragments should be accounted for by experimentally derived data, and vice versa
  - Requires at least 2X fingerprint fragment coverage in 2 domains at any given point in assembly

Pme I: PFGE & Sequence Fragments



Pulse Field Gel Electrophoresis

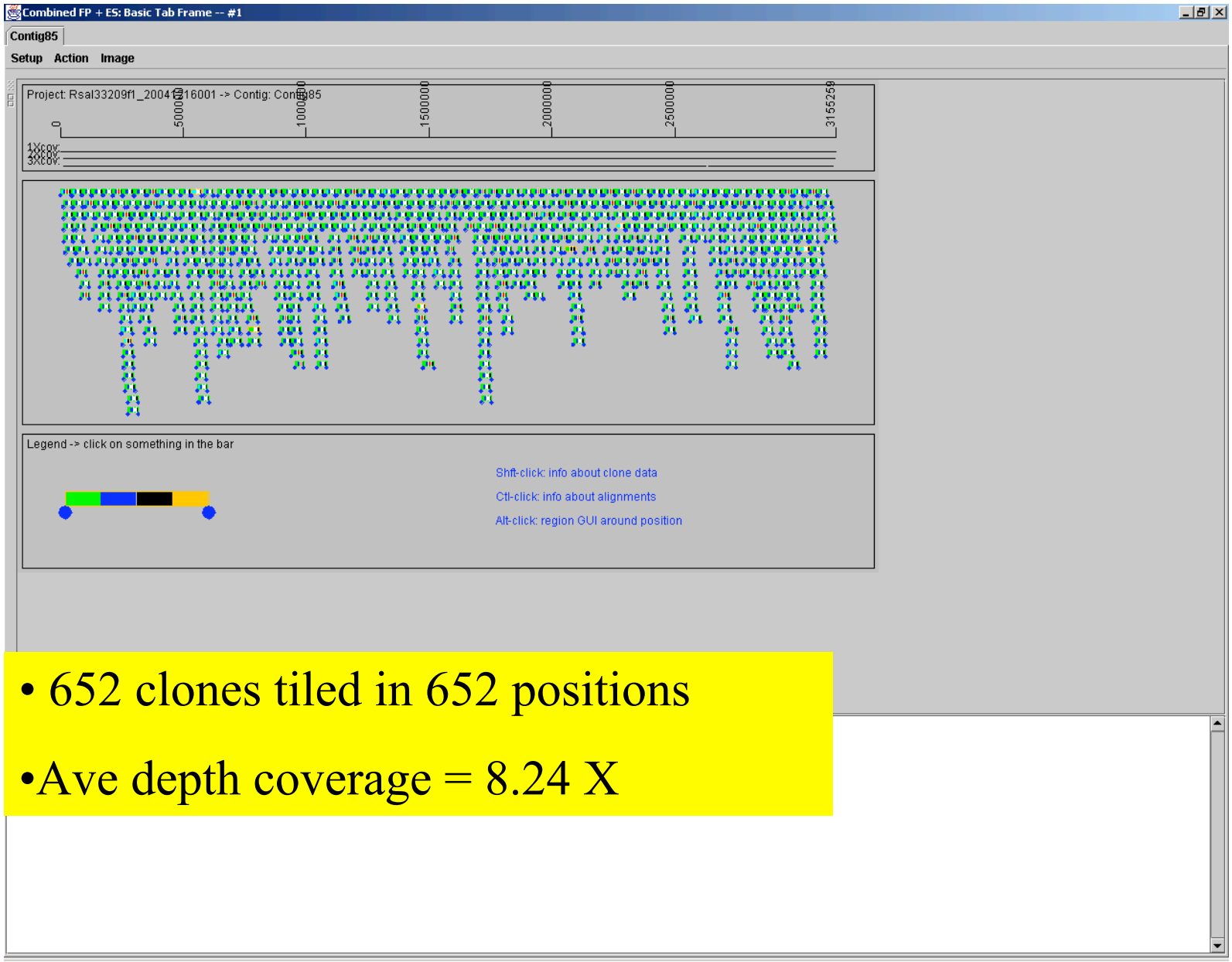


Fosmid End-sequences and FP data

- **Clones = 768 (9.7 X clone coverage)**
  1. clones with X end seq = 728
  2. clones with Y end seq = 725
  3. clones with NO end seq = 20

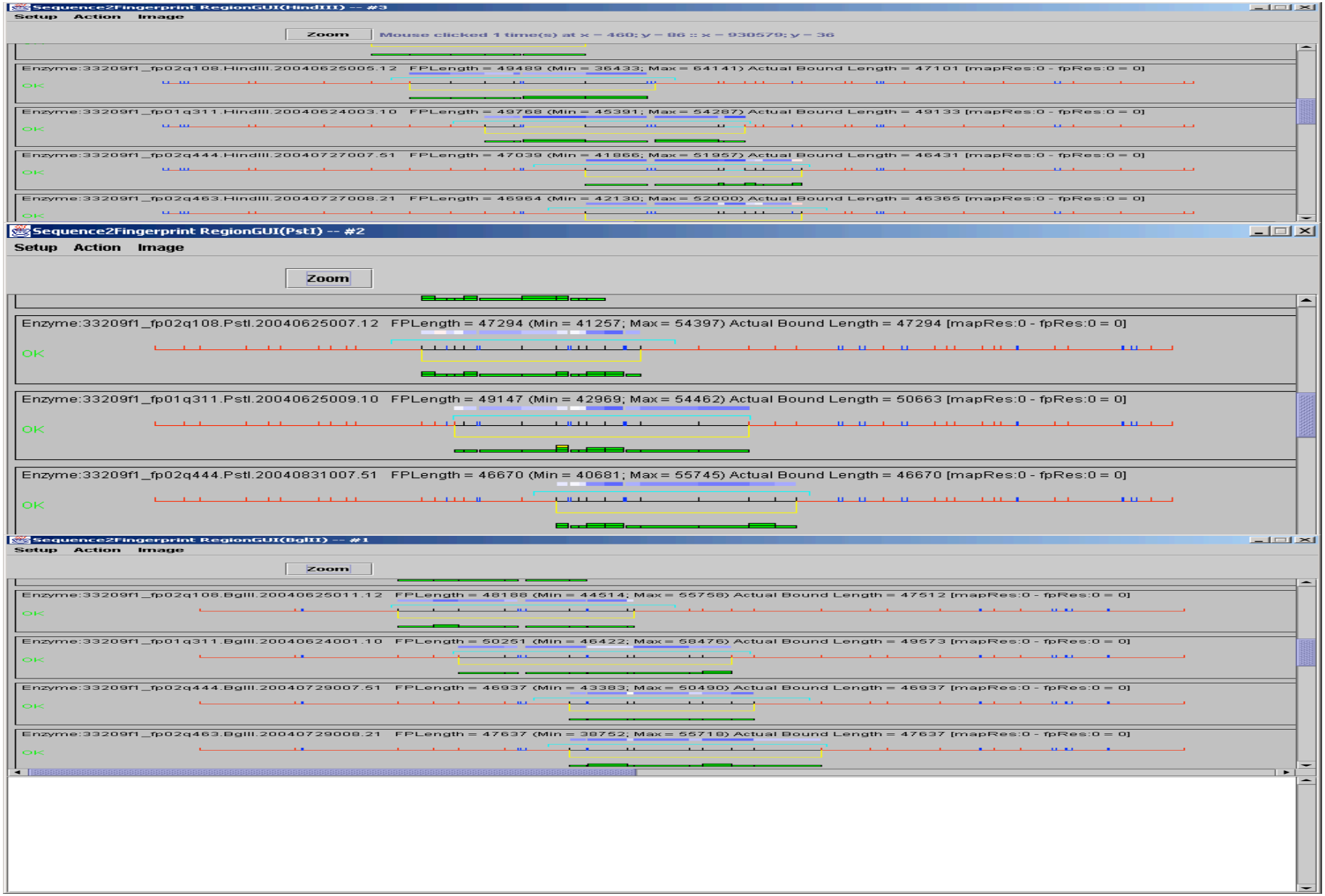
- **Clones with 3 enzymes present = 730**
  1. clones with BglII = 755
  2. clones with PstI = 754
  3. clones with HindIII = 747
  4. clones with (exactly) 0 enzymes present = 4
  5. clones with (exactly) 1 enzymes present = 2
  6. clones with (exactly) 2 enzymes present = 32

Best Full Position Tiling

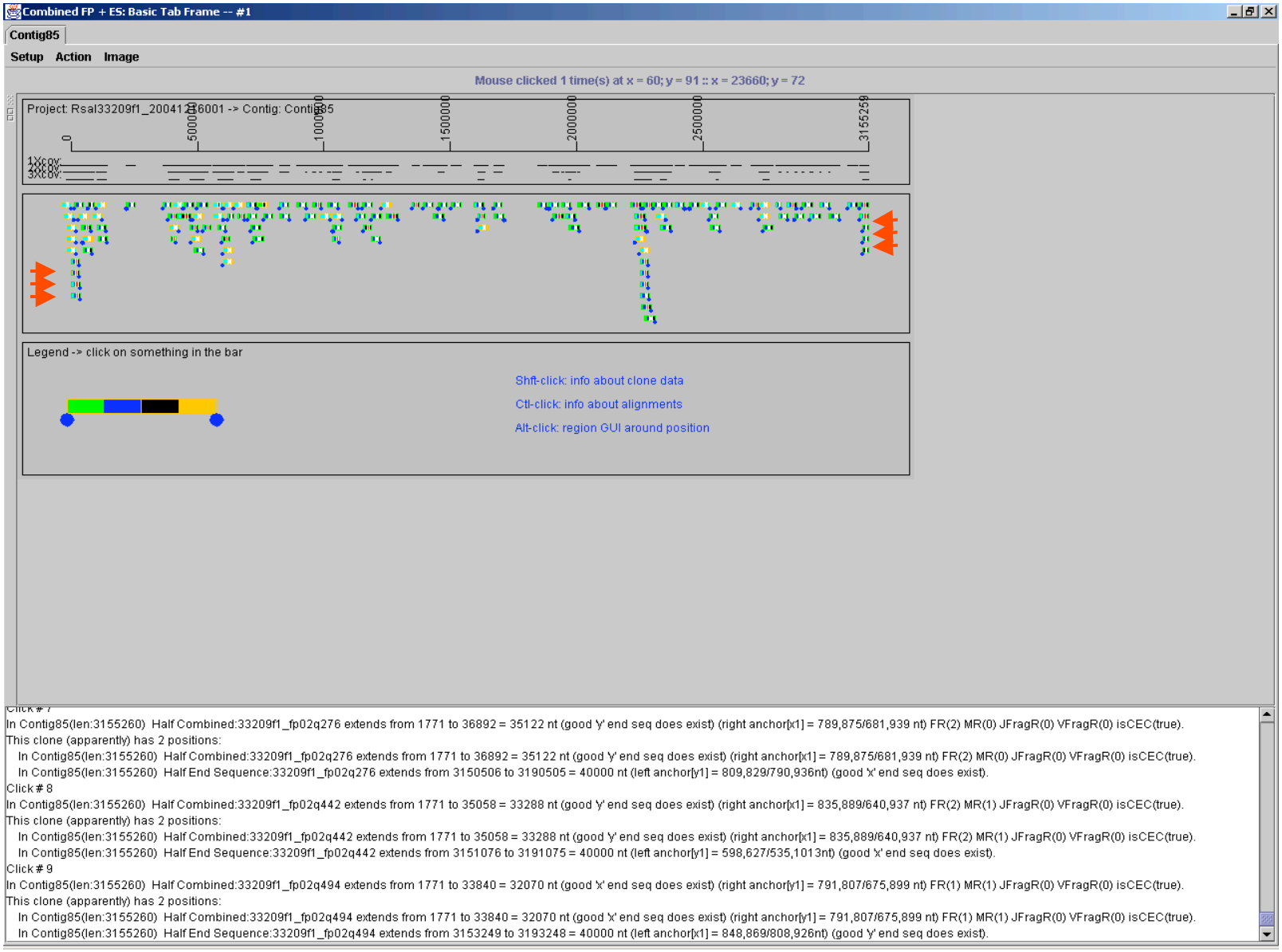


- 652 clones tiled in 652 positions
- Ave depth coverage = 8.24 X

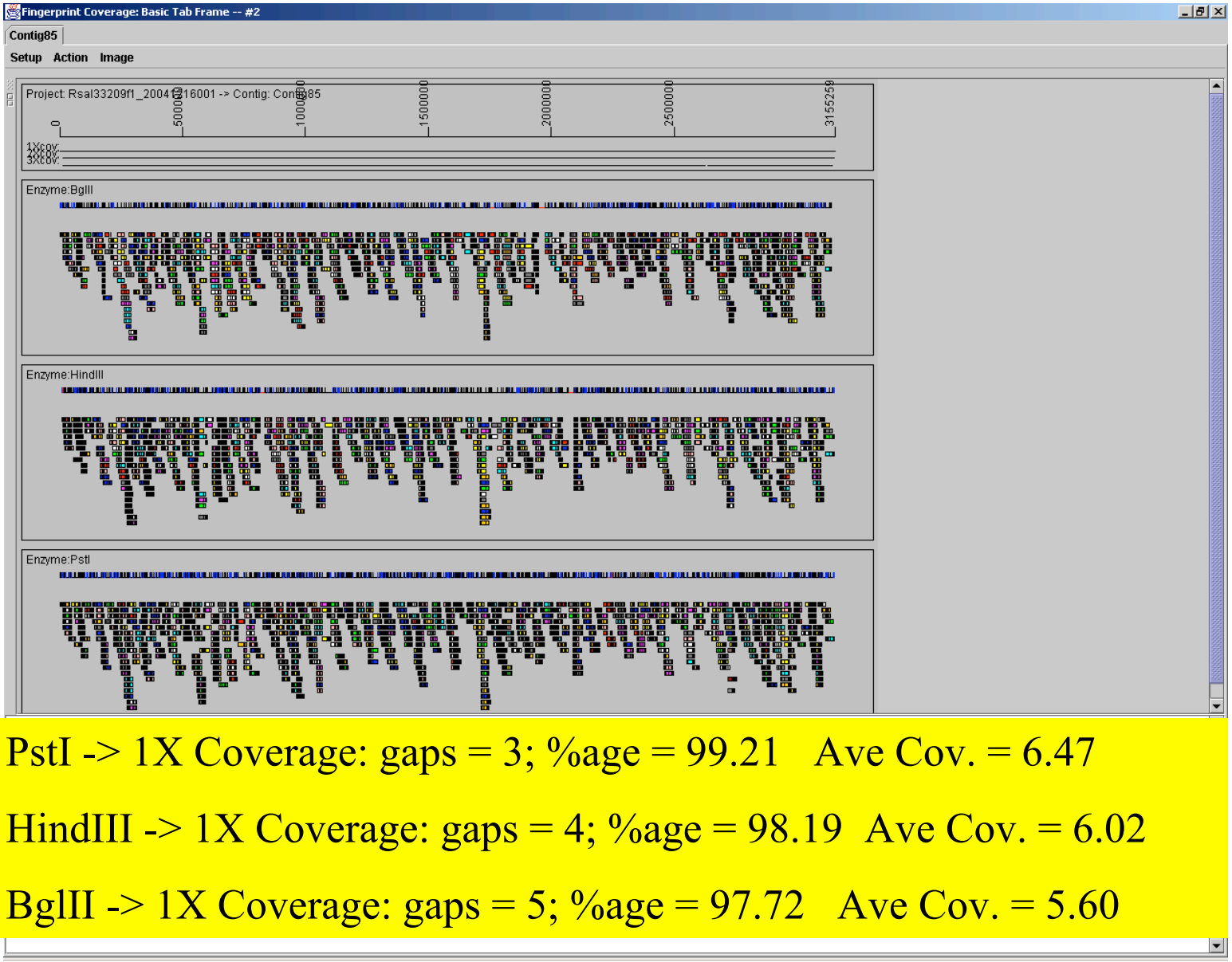
Restriction Fragment Coverage in Multiple Domains Across Clones



Clones Validating Circular Chromosome



Fingerprint Coverage

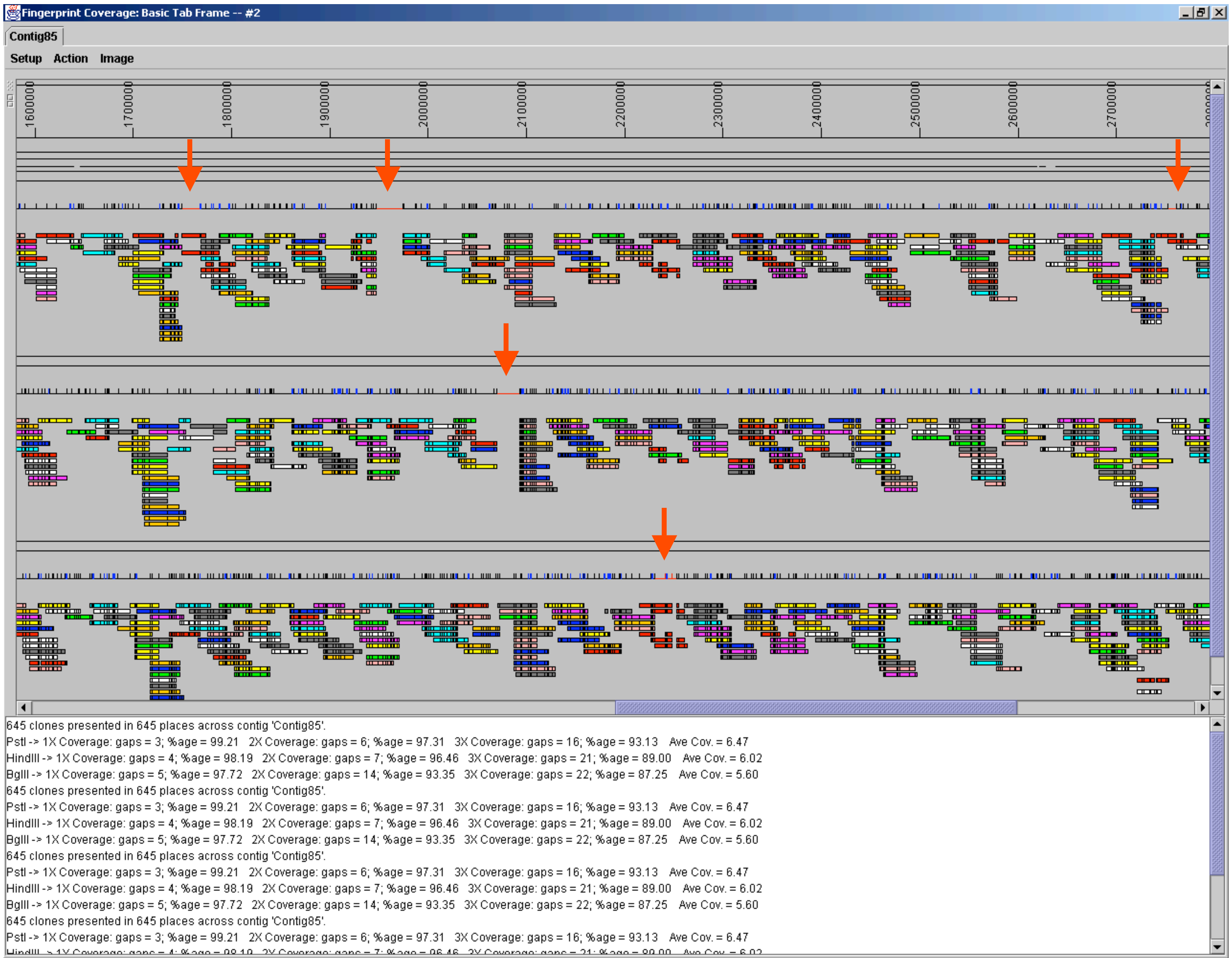


PstI -> 1X Coverage: gaps = 3; %age = 99.21 Ave Cov. = 6.47

HindIII -> 1X Coverage: gaps = 4; %age = 98.19 Ave Cov. = 6.02

BglII -> 1X Coverage: gaps = 5; %age = 97.72 Ave Cov. = 5.60

Non-contiguous Gaps in FP Coverage



Genome Sequencing Highlights

- Single Circular chromosome with **3,155,258** bases
- Two insertional sequence repeat families
  - IS994: 70 copies
  - ISRsa2: 11 copies
- GC content: 56.27%
- Number of ORFs: 3,667
- Sequence with coding potential: 91%
- Average ORF size: 862 bases

Acknowledgements






UWGC Production Staff